Automatic Classification of Academic and Vocational Guidance Questions using Multiclass Neural Network

Omar Zahour[1], El Habib Benlahmar[2], Ahmed Eddaoui[3], Oumaima Hourrane[4]
Laboratory of Information Technology and Modeling
Hassan II University, Faculty of Sciences Ben M'SIK, Casablanca, Morocco

***Abstract*—The educational and professional orientation is an essential phase for each student to succeed in his life and his curriculum. In this context, it is very important to take into account the interests, occupations, skills, and the type of each student's personalities to make the right choice of training and to build a solid professional outline. This article deals with the problematic of educational and vocational orientation and we have developed a model for automatic classification of orientation questions. "E-Orientation Data" is a machine learning method based on John L. Holland's Theory of RIASEC typology that uses a multiclass neural network algorithm. This model allows us to classify the questions of academic and professional orientation according to their four categories, thus allows automatic generation of questions in this area. This model can serve E-Orientation practitioners and researchers for further research as the algorithm gives us good results.**

Keywords—Academic and vocational guidance; multiclass neural network; e-orientation; machine learning; Holland's theory

I. Introduction

The classification of questions is a problem that has already been studied by several researchers in this field, but most of the work is domain-specific or limited to a high-level classification.

Anbuselvan and R.Ahmed [1] proposed an SVM-based method for the same task. The question is first analyzed and numbered, the parts of the speech are labeled, the stop words are deleted, the data is truncated and many features are extracted. The feature selection steps are performed prior to transmitting the data to a carrier vector machine for training. The same treatment is also done for test questions, which can take a long time to get results in real-time.

Marco Pota [2] propose a feature-based method, in which features related to a subset of questions such as keywords, how - all / some words, leading verbs and various other such features were extracted from the texts a classifier.

For Natural Language Processing (NLP) Convolutional neural networks (CNNs) have already been used in some works. Collobert and J.Weston [3] first proposed the idea of a convolutional neural network architecture, which includes lookup tables and hyperbolic hard tangents. Kalchbrenner and P.Blunson [4] proposed a simplified version of Collobert's network, which was used to classify Twitter's questions and opinions. They used the concept of k-max pooling. Yoon Kim [5] developed Kalchbrenner's work to add various machine learning strategies, such as regularization, to improve network performance.

For the time, the question classification has mainly been studied in the context of open-domain TREC (Text REtrieval Conference) questions [6], with smaller recent datasets available in biomedical [7] [8] and education [9]. The TREC corpus of questions from the open-domain is a set of questions associated with a taxonomy developed by Li and Roth [10] that includes 6 types of coarse responses (such as entities, locations and numbers) and 50 fine-grained types (for example, specific types of entities, such as animals or vehicles). While a wide variety of syntactic, semantic and other features and classification methods have been applied to this task, culminating in an almost perfect classification performance [11], recent work has shown that QC methods developed on TREC issues usually fail to transfer to datasets with more complex issues such as those in the biomedical field [7], probably due in part to the simplicity and syntactic regularity of questions and the possibility of simpler term frequency models achieve near-ceiling performance [12].

In this world, the educational and guidance system of each country seeks to help the students or the laureates of higher education institutions and vocational training institutes to make their choice.

According to Ali Boulahcen [13] and through his analysis, he noticed that there is no real process of educational guidance in Morocco, but there is only a summary process in the context, within a few seconds, one decides on the fate of the pupil that based solely on his academic value then translated by a numerical note.

This means that the Moroccan school institution is based at least on selection criteria and not on orientation [13].In this context, our goal is to set up an E-Orientation system that is interested in the automation of the orientation task, thanks to the evolution of information technologies. The realization of this electronic guidance system requires the classification then modeling and integration of user preferences in this system. In this paper, we used the Multi-Class Neural Networks algorithm to classify the different questions according to John L. Holland's RIASEC topology.

This document is organized as follows:

Section 2 provides a literature review of the various theories of educational and vocational guidance, including the theory of John L. Holland, Section 3 deals with the different automatic classification algorithms of the text, Section 4 deals with the experimental evaluation of the classification, and Section 5 covers the results obtained and the conclusion with research perspectives.

II. Holland's Theory and RIASEC Types

A. *Holland's Theory*

The guiding approach is based on theories and studies related to career choice and career development. These include Hoyt's concept of career education, Gardner's theory of multiple intelligences and Holland's typology of professional interests [14]. Holland's theory of vocational choice (1997) [15], is the result of the work of American psychologist and researcher "John Holland (1919-2008)". The results of his research argue that the association of workers to one type of career would be determined by their skills, interests, and personality.

Some activities would be better suited to one type of person than another. It constitutes the theoretical anchoring of our classification model and serves as a basis for many psychometric tools, including the Hexa 3d professional interest's questionnaire.

Although this theory, dating from the mid-1960s is still widely used [16] and has been the subject of numerous studies [17] - [18].

To briefly explain his theory, Holland (1997) [15] formulates several hypotheses according to professional interests that are a mode of expression of personality. Therefore, he considers the choices of orientation as a mode of expression of this personality and distinguishes six types of personality (RIASEC), according to aptitudes, personality traits, values, and beliefs.

Of all the models related to career development, the Holland model has been the subject of the greatest number of analyzes and studies.[19].Among those conducted on the structure of interests across gender and ethnic populations, a number demonstrates the consistency of the arrangement of types and their proximity on a hexagonal and spherical model [18], [20], [21]. This debate focuses more on the geometric regularity of the hexagon and on the correspondence distances between the different types. Vrignaud and Bernaud (1994) validated other things such as the structure of the Holland model in France [22].

Professional activities, as well as work environments, tend to bring together people who share common interests to a certain extent. The choice of a profession or trade is a form of expression of the personality of an individual; it is the theory of vocational interests.

The person-work environment combination is the most widely used method in the world of educational and vocational guidance.

B. *Holland's RIASEC Types*

The theory of vocational choice distinguishes six categories of professional interest (realistic, investigative, artistic, social, enterprising, and conventional) corresponding to different personality profiles. Holland represents them according to a hexagonal model illustrated in Fig. 1 [23].

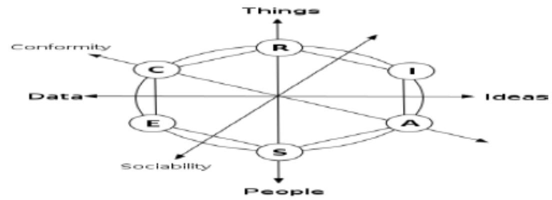

Fig. 1. Representation of Holland's Circular Model (RIASEC).

According to Holland - and many research, they have confirmed the profession or trade chosen by a person which is a form of expression of his personality. Therefore, it is related to the type to which he belongs.

The affiliation of a worker to one of the six types would be determined by his aptitudes, by certain traits of his personality and interests. So, according to Holland, people of the same type would be attracted to the same kind of work. Why? Because these people are similar in their personality, in the fact they pursue similar objectives and have the same physical or psychological dispositions with regard to their work. All persons can be divided into six professional types.

The typology of a person is established by measuring his degree of affinity with each of the six types, to place them in order of importance, of the type that corresponds most to him. For most people, it is mostly the first two or three types of personal classification that determine their way of being and acting in their personal and professional lives. For example, a person whose dominant type is "Investigator" and who has affinities with the "Realist" type; we will say that he has an "IR" profile. To further characterize this person's typology, it is possible to consider the third type which it most closely resembles and to say the case where it is of the "Social" type and is this person has an "IRS" profile?

These types can be combined in all sorts of ways and their combination determines the personality.

THE REALISTIC TYPE – People exercise mainly concrete tasks. With their hands, they know how to coordinate their actions. They use tools; operate appliances, machines, vehicles. Realists have a sense of mechanics which is a concern for precision. Many people practice their profession outdoors rather than indoors. Their work often requires good physical stamina and even athletic abilities. These people are patient, meticulous, consistent, sensible, natural, frank, practical, concrete and simple.

TYPE INVESTIGATOR - Most people of this type have theoretical knowledge which uses it to act. They have specialized information that they use to solve problems. They are observers. Their main competence lies in their understanding of phenomena. They like to be absorbed in their reflections. They like to play with ideas. They value knowledge. These people are critical, curious, anxious to learn, calm, reserved, persevering, tolerant, cautious in their judgments, logical, objective, rigorous and intellectual.

THE ARTISTIC TYPE - People of this type like activities that allow them to express themselves freely based on their perceptions, their sensitivity, and their intuition. They are interested in creative work, visual art, literature, music, and advertising or entertainment. With an independent and unconventional spirit, they are comfortable in situations that are out of the ordinary. They have great sensitivity and a lot of imagination. Although discouraged by methodical and routine tasks, they are nevertheless able to work with discipline. These people are spontaneous, expressive, imaginative, emotional, independent, original, intuitive, passionate, proud, flexible, and disciplined.

THE SOCIAL TYPE - People of this type like to be in contact with others in order to help them, to inform them, to educate them, to entertain them, to treat them or to promote their growth. They are interested in human behavior and care about the quality of their relationships with others. They use their knowledge and their feelings and emotions to act and interact with others. They like to communicate and express themselves easily. These people are attentive to others, cooperative, collaborating, understanding, dedicated, sensitive, friendly, insightful, caring, communicative, and encouraging.

THE ENTREPRENEURIAL TYPE - People of this type like to influence their surroundings. Their decisions make ability; a sense of organization and a particular ability to communicate their enthusiasm to support them in their goals. They know how to sell ideas as much as material goods. They have a sense of organization, planning, initiative, and know-how to carry out their projects. They know how to be bold and efficient. These people are persuasive, energetic, optimistic, audacious, self-confident, ambitious, determined, diplomatic, resourceful, and sociable.

THE CONVENTIONAL TYPE - People of this type have a preference for specific, methodical activities that focused on a predictable outcome. They are concerned about the order and the good material organization of their environment. They prefer to abide by well-established conventions and clear instructions rather than acting in improvisation. They like to calculate, classify, and maintain records or records. They are effective in any job that requires accuracy and ease in routine tasks. These people are loyal, organized, efficient, and respectful of authority, perfectionist, reasonable, conscientious, punctual, discreet, and strict [24].

III. RELATED WORK

Classification systems for the best-performing questions tend to use a rule-based custom template matching [25] [11], or a combination of basic learning approaches. of rules and machine learning [26], to the detriment of model construction time.

Recent research on the methods learned has shown that a large number of CNN variants [27] and LSTM [12] achieve similar precision on the TREC question classification; these models presenting at best small gains compared to simple models term frequency models. These recent developments echo the observations of Roberts and M.Fiszman [7], who have shown that existing methods beyond term frequency models fail to generalize to questions in the medical field.

In the education sector, researchers Godea.A and Nielsen.R [9] collected 1,155 questions in class and classified them into 16 categories. To allow a detailed study of the classification of questions in the scientific field, the process of classifying a text collection is to label each text with one or more predefined classes (Categories). In this process, an algorithm is first designed then it is driven with a set of specific characteristics, for example, word occurrences or even theme distributions in a document. Once trained, the algorithm is used to label new texts, but these are different from the texts used during training. The algorithm is evaluated on the number of classification errors obtained during the learning phase and during the test phase.

When we are training the classification algorithm, the extraction phase of the characteristics is used for learning crucial. These Characteristics extracted from texts that are typically derived from a large vector space. This space is constructed with vector modeling of words using distributional semantics [28].

Data science or statistical algorithms are further classified into multiple machines learning specific algorithmic categories:

- Supervised learning algorithms (label and output known).
- Unsupervised learning algorithms (label and output not known).
- Reinforced learning algorithms (reward-based agent action).
- Semi-supervised learning algorithms (mix of supervised and unsupervised).

These algorithms, in turn, contain multiple sub-algorithms and types (see Table I). For example, a few algorithms fall under the category of parametric, whereas others are non-parametric. In parametric algorithms, information about the population is completely known which not the case with non-parametric algorithms is. Typically, parametric models deal with a finite number of parameters, whereas non-parametric learning models are capable of dealing with an infinite number of parameters. Therefore, the training data grows the complexity of nonparametric models increases. Linear regression, logistic regression, and Support vector machines are examples of parametric algorithms. K-nearest neighbor and decision trees are non-parametric learning algorithms. These algorithms are computationally faster in comparison to their nonparametric companions.

As Table I depicts, the machine learning algorithms are large in number [29].

In this section, we briefly describe various machine-learning algorithms used for forecasting.

A. *Support Vector Machine (SVM)*

SVM Classifiers attempt to partition the data space with the use of linear or non-linear delineations between the different classes. The key in such classifiers is to determine the optimal boundaries between the different classes and use them for the purposes of classification.

TABLE. I. MACHINE LEARNING ALGORITHMS

Supervised Learning	Unsupervised Learning	Reinforcement Learning
Artificial neural network	Artificial neural network	
Bayesian statistics	Association rule learning	Q-learning
Case-based reasoning	Hierarchical clustering	Learning automata
Decision trees	Partitioned clustering	
Learning automata		
Instance-based learning		
Regression analysis		
Linear classifiers Decision trees		
Bayesian networks		
Hidden Markov models		

B. Naïve Bayes Classifier

Naïve Bayes classifier is statistical classifiers, which predict class membership based on probabilities. Naive Bayes classifiers make use of class conditional independence, which makes it computationally faster. Class conditional independence means every attribute in the given class independent of other attributes. Naive Bayes classifier works as follows:

Let us suppose T represents a training set of samples. There are k classes, so class labels would be $C_1, C_2, \ldots\ldots, C_k$. Each record is represented by an n-dimensional vector, $X = \{X_1, X_2, \ldots\ldots, X_n\}$. It represents n measured values of the n attributes $A_1, A_2, \ldots\ldots, A_n$ respectively. Classifier will predict the class of X based on highest a posteriori probability. Thus we find the class that maximizes $(C_i \mid X)$ By Bayes Theorem, we have k :

$$P(C_i / X) = P(X \mid C_i) P(C_i) / P(X) \tag{1}$$

As $P(X)$ has same value for all classes, we can ignore it. Naïve Bayes makes class conditional independence assumption mathematically:

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) \tag{2}$$

The probabilities $(x_1 \mid C_i), (x_2 \mid C_i), \ldots\ldots\ldots\ldots, (x_n \mid C_i)$ are computed from the training set. In (2), the term x_k denotes the value of attribute A_k for the given sample.

C. K-Nearest Neighbors

K-Nearest Neighbor (KNN) is a simple to implement machine learning classifier. The decision is taken on the basis of similarity parameters such as Euclidean distance. The KNN classifier works as follows:

1) Compute k number of nearest neighbors.

2) Determine the distance between the test samples and the training samples by using metrics such as Euclidean distance.

3) Perform sorting on all the training data is on the basis of distances.

4) Decide class labels of k nearest neighbors on the basis of majority vote and assign it as a prediction value of the query record.

D. Multiclass Logistic Regression

Multinomial logistic regression is a form of logistic regression which used to predict a target variable; it has more than two classes. It is a modification of logistic regression using the softmax function instead of the sigmoid function, and the cross-entropy loss function. The softmax function squashes all values to the range [0,1] and the sum of the elements is 1.

$$soft\max(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{3}$$

Cross entropy is a measure of how different two probability distributions are near to each other. If p and q are discrete we have:

$$H(p,q) = -\sum_x p(x)\, Log\, q(x) \tag{4}$$

This function has a range of [0, inf], it is equal to 0 when p=q and infinity then p is very small compared to q or vice versa. For example x, the class scores are given by vector z=Wx+b, where W is a C×M matrix and b is a length C vector of biases. We define the label y as a one-hot vector equal to 1 for the correct class c and 0 everywhere else. The loss for a training example x with predicted class distribution y and correct class c will be:

$$\hat{y} = softmax(z) \tag{5}$$

$$\begin{aligned} loss &= H(y, \hat{y}) \\ &= -\sum_i y_i\, Log\, \hat{y}_i \\ &= -Log\, \hat{y}_c \end{aligned} \tag{6}$$

As in the binary case, the loss value is exactly the negative log probability of a single example x having true class label c. Thus, minimizing the sum of the loss over our training example is equivalent to maximizing the log-likelihood. We can learn the model parameters W and b by performing gradient descent on the loss function with respect to these parameters. There are two common methods to perform multi-class classification using the binary classification logistic regression algorithm: one-vs-all and one-vs-one. In one-vs-all, we train C separate binary classifier for each class and run all those classifiers on any new example x, we want to predict and take the class with the maximum score. In one-vs-one, we train C to choose 2 classifiers = C(C-1)/2 one for each possible pair of class and choose the class with maximum votes while predicting for a new example.

E. Multiclass Neural Network

A neural network is a set of interconnected layers. The inputs are the first layer and are connected to an output layer by an acyclic graph composed of weighted edges and nodes. We can insert multiple hidden layers between the input and output layers. Most predictive tasks can be accomplished easily with one or more hidden layers. However, Deep Neural Networks (DNNs) [30], [31] with many layers can be very effective for complex tasks such as image recognition or speech. Successive layers are used to model increasing levels of semantic depth. The relationship between inputs and outputs is learned during the formation of the neural network on the input data. The chart direction passes inputs to the hidden layer and the output layer. All the nodes of a layer are connected by the weighted edges to the nodes of the next layer.

To calculate the network output for a particular input, a value is calculated at each node of the masked layers and the output layer. The value is defined by calculating the weighted sum of the values of the nodes of the previous layer. An activation function is then applied to this weighted sum.

We use a multiclass neural network module to predict a multi-valued target knowing that neural networks of this type could be used in complex computer vision tasks, such as recognition of numbers or letters, classification of documents, of text (Questions ...) and also for pattern recognition. In this sense classification using neural networks is a supervised learning method. It, therefore, requires a tagged data set comprising a label column.

IV. Experimental Evaluation and Results

A. Proposed Method

Our proposed system is based on the multi-class neural network algorithm which follows supervised learning. The goal of this is to discover an underlying structure of the data. This algorithm requires a tagged dataset. The data set on orientation questions "E-Orientation Data" is divided into two series, such as training data and test data. The classification performed by the algorithm used in our model is based on the knowledge acquired by the learning data during the learning process.

Our dataset was collected from the RIASEC test based on Holland's theory [32], [33], [34], it contains two columns namely:

Question: It contains questions and statements that measure either the occupations or the activities or abilities or the personality of the users.

Categories: we have four classes (labels) of categories namely:

1) Activity
2) Occupations
3) Abilities
4) Personality

In our research work on Guidance Classification, we used the Azure Machine Learning Studio [35] tool which is a collaborative drag-and-drop tool that we can use to create, test, and deploy predictive analytics solutions on our data. Machine Learning Studio publishes templates as a web of services that can be easily consumed by custom applications. Machine Learning Studio is the meeting place of data science, predictive analytics, cloud resources, and our data.

B. Experiment Steps

The experimental steps described and illustrated in Fig. 4. They are explained below:

1) Importing the Dataset: We import our dataset entitled "E-Orientation Data" that we collected from several websites from our local disk on Azure ML Studio to be used for the experiment and Category names that we have been used as a class tag or attribute to predict.

2) Preprocessing and Preparing the Dataset: The dummy column headers have been replaced by meaningful column names by using the metadata editor. Also, missing values have been cleared by deleting the entire line containing the missing value.

3) Feature engineering: After the processing phase of the dataset, we will use the feature hashing module to convert the raw text of the questions into integers; and use the integer values as input entities of the model. Fig. 3 represents our model.

4) Split the Data and Parameter Settings: We have divided the data of "E-Orientation Data" as 70% of the data for training and 30% for the test. Then we applied the Multiclass Neural Networks algorithm with the default settings for model formation. The parameters have been set by using the "Tune model hyperparameters".

5) The Model: We used the Multiclass neural network algorithm. A neural network is a set of layers that are interconnected. The first layer is the inputs which are connected to an output layer by an acyclic graph; it is comprised of weighted edges and nodes. Multiple hidden layers are present between the input and output layers. The relationship between inputs and outputs is obtained for training input data of the neural network. All the nodes in a layer are associated with the weighted edges to the nodes in the successive layer.

6) Score and evaluate the Model: The Evaluate model also visualizes the results through the confusion matrix.

C. The Results

We used the Multiclass Neural Network algorithm to classify the category of academic and professional guidance questions according to Holland's RIASEC typology. The graphic representation of the Multiclass classification is given in Fig. 2.

Fig. 4 shows the overall implementation of our model.

The classification is carried out as shown in Fig. 5.

Prediction accuracy for each class is shown in Fig. 6.

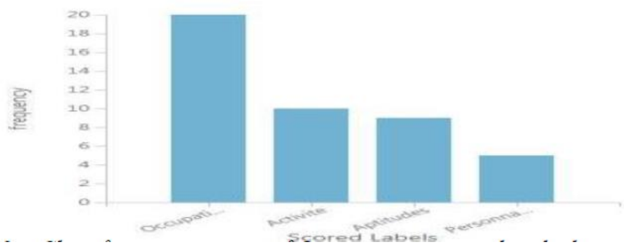

Fig. 2. Classification Category of Question Represented in the Histogram.

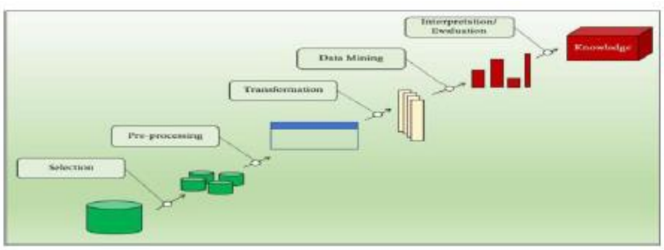

Fig. 3. Schema of Model.

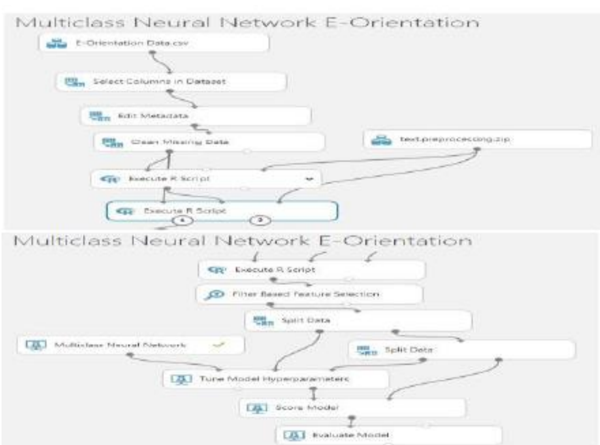

Fig. 4. Classification using Multiclass Neural Network.

Scored Probabilities for Class "Activite"	Scored Probabilities for Class "Aptitudes"	Scored Probabilities for Class "Occupations"	Scored Probabilities for Class "Personnalite"	Scored Labels
0.208896	0.014198	0.644117	0.10673	Occupations
0.000097	0.999752	0.001055	0.000814	Aptitudes
0.001762	0.016079	0.999992	0.002484	Occupations
0.000088	0.999334	0.006043	0.00104	Aptitudes
0.004054	0.011152	0.018266	0.997337	Personnalite
0.949081	0.024299	0.003832	0.060052	Activite

Fig. 5. Snapshot of the Classification.

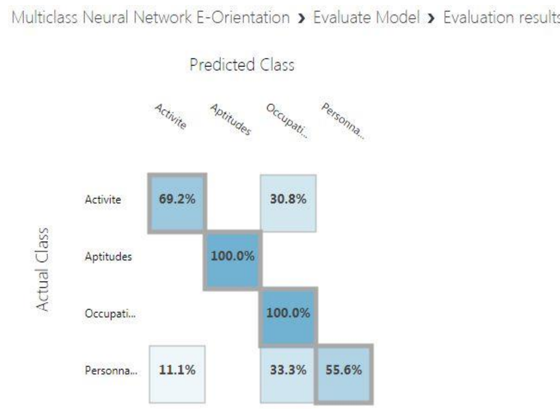

Fig. 6. Confusion Matrix.

V. CONCLUSION

In this paper, we defined the different Machine Learning algorithms used for text classification. We conclude that multiclass neural networks perform better than other algorithms of machine learning.

The Multiclass Neural Network algorithm used in our classification model of academic and professional orientation questions by category are implemented by using Azure Machine Learning Studio. In fact, we found that the supervised method gives very good precision. This method can also be used to automatically generate academic and vocational orientation questionnaires by knowing in advance the class of these new questions proposed, and we can see this research question as a perspective. This model can also help the researchers of e-Orientation in the development process in this area.

As future work, we focus on using social network analysis, for example, using Twitter sentiment analysis as a feature to determine the class of questions and the interests of students and faculties of educational institutions.

VI. Conflict of Interest

The authors declare no conflict of interest.

References

[1] Sangodiah, R. Ahmad, and W. F. W. Ahmad, "A review in feature extraction approach in question classification using Support Vector Machine," in Proceedings - 4th IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2014, 2014.

[2] G. D. Pietro Marco Pota, Angela Fuggi, Massimo Esposito, "Extracting Compact Sets of Features for Question Classification in Cognitive Systems: A Comparative Study," in 2015 ,10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, pp. 551–556.

[3] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," J. Mach. Learn. Res., 2011.

[4] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference, 2014.

[5] Y. Kim, "Convolutional neural networks for sentence classification," in EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2014.

[6] E. M. Voorhees and D. M. Tice, "Building a question answering test collection," SIGIR Forum (ACM Spec. Interes. Gr. Inf. Retrieval), 2000.

[7] K. Roberts, H. Kilicoglu, M. Fiszman, and D. Demner-Fushman, "Automatically classifying question types for consumer health questions," AMIA Annu. Symp. Proc., 2014.

[8] M. Wasim, M. N. Asim, M. U. Ghani Khan, and W. Mahmood, "Multi-label biomedical question classification for lexical answer type prediction," J. Biomed. Inform., 2019.

[9] A. Godea and R. Nielsen, "Annotating educational questions for student response analysis," in LREC 2018 - 11th International Conference on Language Resources and Evaluation, 2019.

[10] X. Li and D. Roth, "Learning question classifiers," 2002.

[11] H. T. Madabushi and M. Lee, "High accuracy rule-based question classification using question syntax and semantics," in COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers, 2016.

[12] W. Xia, W. Zhu, B. Liao, M. Chen, L. Cai, and L. Huang, "Novel architecture for long short-term memory used in question classification," Neurocomputing, 2018.

[13] A. Boulahcen, "Le processus d'orientation scolaire au MarocA Sociological Analysis of the Careers Advice Process in Moroccan SchoolsUn análisis sociológico del proceso de orientación escolar en Marruecos," Rev. Int. d'éducation Sèvres, 2005.

[14] G. . & M. . D. Dupont.P, "Guide for Educational and Vocational Information and Guidance: Elementary and Secondary Cycle Three, Youth Sector," Prov. Support Gr. Youth Orientat. approach to Sch. Univ. Sherbrook, Sherbrook, Canada., 2002.

[15] J. L. Holland, "Making vocational choices: A theory of vocational personalities and work environments," Englewood Cliffs NJ PrenticeHall, 1997.

[16] M. M. Nauta, "The Development, Evolution, and Status of Holland's Theory of Vocational Personalities: Reflections and Future Directions for Counseling Psychology," J. Couns. Psychol., 2010.

[17] R. Du Toit and G. P. De Bruin, "The structural validity of Holland's R-I-A-S-E-C model of vocational personality types for young black South African men and women," J. Career Assess., 2002.

[18] D. Guglielmi, F. Fraccaroli, and M. L. Pombeni, "Les intérêts professionnels selon le modèle hexagonal de Holland," L'orientation Sc. Prof., 2004.

[19] A. R. Spokane, E. J. Luchetta, and M. H. Richwine, "Holland's theory of personality in work enviornments," in Career choice and development, 2002.

[20] A. R. Spokane and M. C. Cruza-Guet, "Holland's Theory of Vocational Personalities in Work Environments," in Career development and counseling: Putting theory and research to work, 2005.

[21] Tetreau.B, "The rise of a psychology of professional interests," Carrierology, vol. 10, no. 1, pp. 77–118, 2005.

[22] V. P. & Bernaud.J, "The interests of the French People are they hexagonal?" 1. Elements for the validation of the model of the interests of Holland (RIASEC) in France," Quest. Orientat., vol. 49, no. 1, pp. 17–39, 1994.

[23] "Holland's Codes," https://en.wikipedia.org/wiki/Holland_Codes. [Online]. Available: https://en.wikipedia.org/wiki/Holland_Codes.

[24] "Dictionnaire septembre des métiers et professions; suivi du guide Cléo, des clés pour s'orienter.," Septembre editeur S.E.N.C. pp. 256–257, 2005.

[25] A. Lally et al., "Question analysis: How Watson reads a clue," IBM Journal of Research and Development. 2012.

[26] J. Silva, L. Coheur, A. C. Mendes, and A. Wichert, "From symbolic to sub-symbolic information in question classification," Artif. Intell. Rev., 2011.

[27] T. Lei, Z. Shi, D. Liu, L. Yang, and F. Zhu, "A novel CNN-based method for question classification in intelligent question answering," in ACAI 2018 Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence, p. 54.

[28] Z. S. Harris, "Distributional Structure," WORD, 1954.

[29] P. Kashyap and P. Kashyap, "Machine Learning Algorithms and Their Relationship with Modern Technologies," in Machine Learning for Decision Makers, 2017.

[30] Hourrane, Oumaima, and El Habib Benlahmar. "Survey of plagiarism detection approaches and big data techniques related to plagiarism candidate retrieval." Proceedings of the 2nd International Conference on Big Data, Cloud and Applications. ACM, 2017.

[31] Hourrane, Oumaima, et al. "Using Deep Learning Word Embeddings for Citations Similarity in Academic Papers." International Conference on Big Data, Cloud and Applications. Springer, Cham, 2018.

[32] Omar Zahour,El Habib Benlahmar,Ahmed Eddaoui " E-orientation : entre prescription des théories et prise de décision " Conference TIM'16 ,2016.

[33] Omar Zahour,El Habib Benlahmar,Ahmed Eddaoui " E-Orientation: Between prescription of theories and decision-making " Conference SITA'16 , October 2016.

[34] Omar Zahour,El Habib Benlahmar,Ahmed Eddaoui "E-orientation : Vers une modélisation des facteurs d'orientation scolaire " Conference TIM'18 , 2018.

[35] Team, Azureml,Dorard, Louis,Reid, Mark D,Martin, Francisco J, "AzureML: Anatomy of a machine learning servicein" JMLR: Workshop and Conference Proceedings 50:1–13, 2016.